Hallásmodellre alapozott optimális jeltisztítási eljárás alkalmazásával szerzett tapasztalatok

FÖLDVÁRI RUDOLF
Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési Tanszék
GYIMESI LÁSZLÓ
Győri Széchenyi István Egyetem, Digitális Elektronikai Laboratórium, gyimesi@sze.hu

Kulcsszavak: akusztika, hangosság érzet, kritikus sávszélesség, szűrők, transzformációk

A Földvári-féle hallásmodellben használt általánosított amplitúdó- és frekvencia-transzformáció (GAFT – Generalized Amplitude and Frequency Transformation) ismertetése után bizonyítások nélkül felsoroljuk annak tulajdonságait. Bemutatjuk az optimális jeltisztítási eljárás blokkvázlatát, és a háttérzaj becslésének módszerét. Számos DEMO-t adunk, melyekhez rövid értékeléseket mellékelünk, továbbá közöljük az eredeti és a tisztított wav fájlok, valamint a PC-n futtatható exe fálj elérhetőségét.

1. Hallásmodellezés

A hallás és a hallásmechanizmussal kapcsolatos kérdések évezredek óta foglalkoztatják az emberiséget [6]. Egészen a 20-ik század végéig rendkívül ellentmondásos elméletek kerültek napvilágra [1, 2, 3, 4 és 9]. A 60-as évek közepétől azonban már nyilvánvalóvá vált, hogy lineáris transzformációk segítségével nem magyarázható, ennek egyik legalapvetőbb tulajdonsága, az idő- és frekvenciatartományban való igen jó felbontóképessége [7].

Fizikai-érzeti leképzés

A hangforrás fizikailag mérhető mennyiségei, az intenzitás, frekvencia, időtartam, hangszín és irány, pszichológiai érzeteket váltanak ki a megfigyelőben. A fizikai inger az érzékszerven, idegi vezetésen és agyműködésen keresztül alakítja ki az érzetet. Az egyes ingerek, azaz a fizikai összetevők és az érzet, azaz pszichológiai összetevők között nincs kölcsönösen egyértelmű kapcsolat, összefüggéseik rendkívül bonyolultak [9].

Az érzeti oldal egyes mennyiségeit módunkban áll számszerűleg megismerni, ha a méréshez előzetesen sikerül skálát felállítani. Ez minden pszichológiai kísérlet alapja, és egyben a legnehezebb lépése is. Az érzékelt hangosság és hangmagasság elsősorban a hang intenzitásától és frekvenciájától függ, de a színkép, időtartam és irány is befolyásolja hangosság és hangmagasság érzetünket.

Hangosság

Barkhausen (1927) vezette be a phon fogalmát, amely definíciószerűen a dB értékekkel egyezik meg 1000 Hz-en, más frekvenciákon pedig a Fletcher-Munson görbékről olvasható le [5].

Hangmagasság

A hangmagasság érzete a frekvenciával logaritmikusan növekszik és a legjellegzetesebb intervallum az oktáv. A hangmagasság érzet rendkívül erősen függ attól, hogy a hangokat egyszerre, vagy egymásután szólaltatjuk meg.

Kétféle hangmagasság érzetünk működik, egy melodikus és egy harmonikus. Hangmagasságnak a melodikus skálát fogadjuk el, ugyanis az egyszerre megszólaltatott hangok harmónia érzetet váltanak ki, melynek nincs közvetlen köze a hangmagasság érzethez [9]. A melodikus hangmagasság skála kísérletileg meghatározott összefüggés a frekvencia és a hangmagasság érzet között. Az érzeti skála sem lineárisan, sem logaritmikusan nem függ a frekvenciától.

Az alaphang felismerése

A természetben tisztán szinuszos hang alig fordul elő. A tiszta szinuszhoz a furulya, fuvola és az orgona hangja áll a legközelebb. Az alaphangon kívül annak az egészszámú többszörösei is jelen vannak. Az alaphang érzékelésével áttekinthetetlenül sok irodalom foglalkozik, melyek részben ellentmondóak. Az akusztikai Ohm törvény (1843) szerint a hang magassága a legalacsonyabb Fourier összetevő értékével azonos. Később Helmholtz is csatlakozott Ohm elképzeléséhez [1, 2]. Az alaphang azonban akkor is tisztán hallható, ha a megszólaltatott hang a legmélyebb összetevőt nem tartalmazza.

A jelenség legegyszerűbb, de a valóságnak egyáltalán nem megfelelő magyarázata, a közép- és belsőfül nonlinearitására való hivatkozás, mely szerint a hiányzó alaphang torzítás eredményeképpen keletkező különbségi hang. A legmeggyőzőbb kísérlet, mely bizonyítja, hogy nem „különbségi hang” jön létre, rendkívül egyszerű. Egy $2f_o$ frekvenciájú hangot az egyik, $3f_o$ frekvenciájút pedig a másik fülben megszólaltatva, az f_o frekvenciájú virtuális hang változatlanul hallható, pedig ez esetben az egyik alaphártyát csak az egyik, a másik alaphártyát pedig csak a másik hanggal ingereltük. Ezzel bizonyítható, hogy a virtuális hang agyi eredetű, és semmi köze sincs a különbségi hanghoz, mely nem jöhet létre kétfülű (dichotikus) gerjesztés esetén.

2. Kritikus sávok, fázishatár-frekvencia és két frekvencia-határ fogalma

Kritikus sávok értelmezése

A hangosságérzetünk függ az ingerlő jel sávszélességétől. Akár sok szinuszos hanggal, akár zajjal gerjesztjük a fület, a sáv szélesedésével csak a fizikai hangintenzitás változásával halljuk a jelet hangosabbnak. Ha azonban ez a sávszélesség egy határértéket túllép, akkor megváltozik a helyzet, erőteljesebben növekszik a hangosság érzete, mint ahogy azt az ingerlő hatás növekedése indokolná.

Gondos vizsgálatokkal sikerült tisztázni ezeknek az összefoglaló képességgel rendelkező frekvenciasávoknak az értékét, melyeket kritikus sávoknak nevezünk [6].

Fázishatár-frekvencia értelmezése

Az emberi hallás nemcsak a hangosság érzékelése során mutat egy kritikus sávon belül más tulajdonságot, mint szélessávban, hanem egy adott frekvencia környezetében a fázisra is érzékeny. Ha a frekvenciasávot szélesítjük, akkor egy határ után ez az érzékenység megszűnik, és már nem tudunk az amplitúdó- modulált és a frekvenciamodulált jel között különbséget tenni [6].

Feltűnő megegyezés, hogy ezen a területen belül a különböző frekvenciák hangingere az energiával arányosan okoz hangosságérzetet, azaz megegyezik a kritikus sávokkal. Érdekes, hogy míg a hallásküszöb görbe alakulásában az egyes egyedek között nagy szórás mutatkozik, ezeknek az összefoglaló tulajdonságú sávoknak az értéke egyénektől függetlenül egyetemlegesen érvényes adatnak tűnik.

Két tiszta hang érzékelése

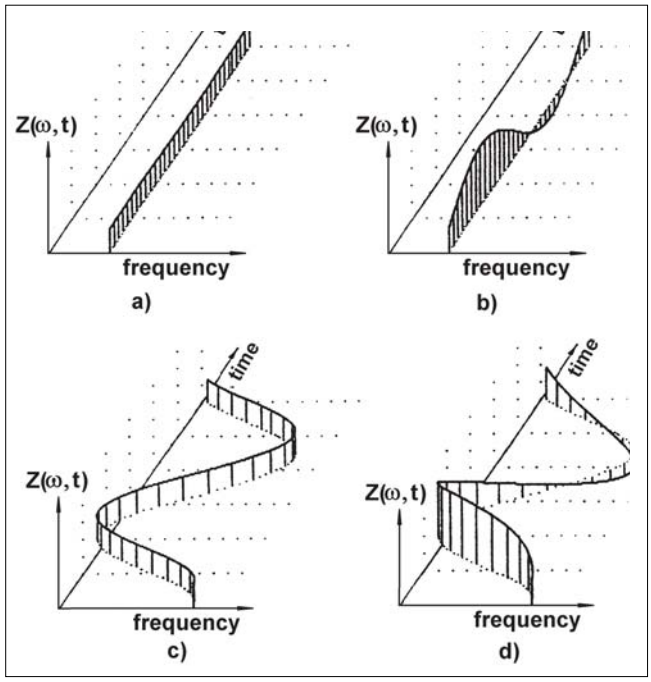
Ha két tiszta hang egyszerre szólal meg, és frekvenciájuk azonos, akkor a hangmagasság nem változik, de ha kissé eltérnek egymástól, akkor előbb lebegést, majd ha még jobban különböznek, érdességet észlelünk. Nem két különböző frekvenciájú hangot, hanem a két frekvencia számtani átlagának megfelelő egyetlen hangmagasságú hangot hallunk. További távolodáskor az érdes, kellemetlen hang egyszer csak két külön hangra hasad szét. Ezt az értéket megkülönböztetési frekvenciatávolságnak, vagy két hang érzeti határnak nevezzük. Nagyjából a kritikus sáv távolságában megszűnik az érdességi megítélés, és ekkor hallunk egymás mellett két „sima“, zavartalan hangot [6].

3. Általánosított amplitúdó és frekvencia transzformáció

Ha egy x(t) időfüggvény Fourier-transzformálható, továbbá nem tartalmaz egyen komponenst, akkor létezik a Hilbert párja, és ezt jelöljük y(t)-vel. Felhasználva x(t) és y(t) időfüggvényeket bevezethetjük a következő transzformációt:

$$A(t) = x(t)\left[\cos\left(\Phi_0 + \int_{t_0}^{t}\Omega(\tau)d\tau\right)\right]^{-1}$$
$$(t) = \frac{d}{dt}\Omega(t) = \frac{x(t)y'(t) - x'(t)y(t)}{x^2(t) + y^2(t)} \quad , \qquad (1)$$

ahol A(t)-t (mely negatív is lehet) általánosított pillanatnyi amplitúdónak, Ω(t)-t pillanatnyi frekvenciának nevezzük. Ezt a függvénytranszformációt Z(ω,t)-vel jelöljük, és **GAFT**-nak (Generalized Amplitude Frequency Transformation) hívjuk [10]. Az (1) egyenlettel adott pillanatnyi paraméterek úgy tekinthetők, mint kölcsönösen független és ideális AM és FM demodulátorokat megvalósító transzformáltak. Az elmondottakat az *1. ábra* szemlélteti.



1. ábra a) modulálatlan vivő b) amplitúdó moduláció c) frekvencia moduláció d) együttes amplitúdó és frekvencia moduláció

A pillanatnyi paraméterek (GAFT) tulajdonságai

• A pillanatnyi paraméterek az idő-frekvencia sík felett egy görbét határoznak meg (1. ábra). Az (1) egyenletből látható, hogy a kapcsolat nemlineáris, a pillanatnyi paraméterekre a szuperpozíció elve nem érvényes.

• A GAFT a geometriai értelemben hasonló jeleket hasonló függvényekbe képezi le.

• A GAFT invariáns az időeltolással szemben.

• A jel pillanatnyi teljesítménye $A^2(t)$.

• Az általánosított amplitúdó és pillanatnyi frekvencia tartója azonos a jel időtartománybeli tartójával.

• A GAFT tetszőlegesen sokszor ismételhető, ha az A(t) és Ω(t) jelek DC komponenseit leválasztjuk. Ilyenkor hasonló tulajdonságú függvényeket kapunk, mint az x(t).

• Ha egy jel periodikus, akkor A(t) és Ω(t) szintén periodikus.

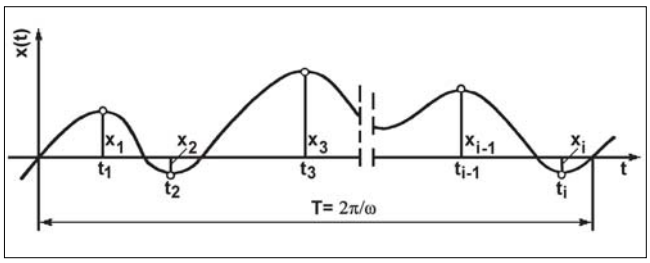
Az úgynevezett bizonytalansági reláció közvetlenül nem értelmezhető a GAFT esetében. A pillanatnyi paraméterek által meghatározott A(t), Ω(t), mint az a (1) egyenletből látható, csak az időtől függ, zérus „szórású". Az illesztett mintavételezés hasonló tulajdonságokkal rendelkezik.

4. Illesztett mintavételezés

Egy sávkorlátos periodikus jel mindig felírható a következő alakban (2):

$$x(t) = \sum_{n=n_L}^{n_H} \left(a_n \sin n\omega t + b_n \cos n\omega t\right) = \sum_{n=n_L}^{n_H} c_n \sin\left(n\omega t + \Phi_n\right)$$

Legyen x(t) a (2)-nek megfelelő alakú, és a (0,T) intervallumban vegyünk mintát a jel helyi szélsőértékeinél a *2. ábrának* megfelelően. Bizonyítható, hogy a $\{x_1,t_1, x_2, t_2,...., x_i, t_i,\}$ halmaz egyértelműen meghatározza x(t)-t, ha $\omega_H<2\omega_L$, azaz ha x(t) komponensei egy oktávnál szűkebb sávba esnek [8]. Ez a feltétel esetünkben teljesül, ugyanis a Zwicker-szűrők kb. terc szélességűek.

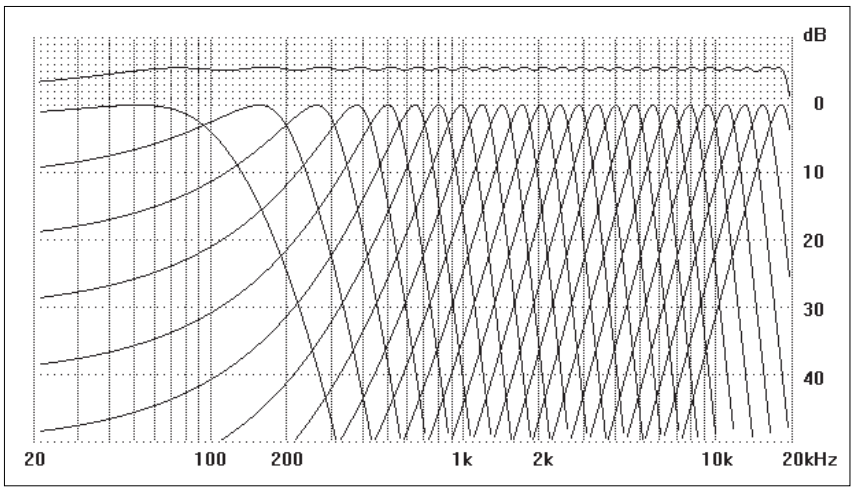


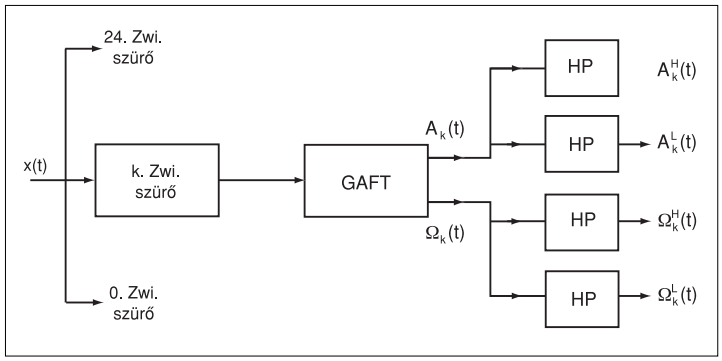
2. ábra
Periodikus, véges sávszélességű jel helyi szélsőértékei

5. A hallásmodell felépítése és tulajdonságai

Az emberi hallás alapvető tulajdonságaira pontos magyarázat adható a Zwicker-féle szűrősor kimenetein alkalmazott GAFT (vagy illesztett mintavételezés) felhasználásával *(3. ábra)*.

3. ábra
Zwicker-féle szűrők (0-24) karakterisztikái és eredőjük





4. ábra A hallásmodell tömbvázlata

Az általunk javasolt legegyszerűbb hallásmodell blokkvázlata a *4. ábrán* látható. A váltószűrők feladata a lassan változó komponens és a kváziperiodikusan változó komponensek szétválasztása. További kiegészítésekre attól függően van szükség, hogy a modell felhasználásával milyen feladatot kívánunk megoldani.

A hallásmodell alapvető tulajdonságai:

• A 3. ábrán jól látható, hogy a szűrőbank eredő karakterisztikája tökéletesen meghatározza a bemeneti x(t) jelet. Az ingadozás kisebb, mint 0.5 dB, a késleltetés kb. 10 ms. Az összegezett jel, még zene esetén sem különböztethető meg az eredetitől.

• Tekintettel arra, hogy mind a GAFT, mind az illesztett mintavétel egzakt transzformáció, a $A_k(t)$ és $\Omega_k(t)$ jelek, illetve a váltószűrők kimeneti jelei, egyértelműen meghatározzák az x(t) jelet. A kapcsolat egyértelmű, de a nemlineáris transzformáció miatt rendkívül bonyolult. Minden x(t)-hez különböző válaszfüggvények tartoznak, és természetesen a lineáris szuperpozíció nem érvényes.

• Az $A_k(t)$ jelek felhasználásával meghatározható az x(t) jel által okozott hangosság érzet. A szűrőbankot az „igazi" hangosság méréséhez használják, hiszen hangosság érzetünk döntően függ attól, hogy az inger komponensei melyik frekvenciatartományba esnek.

• Hosszan megszólaltatott tiszta hang esetén a modell szinte tetszőleges felbontást valósít meg a frekvenciatartományban. Triviális, hogy végtelen hosszú szinuszos jel esetén a Zwicker-féle szűrősor a jelre nincs hatással, kimenetein a pillanatnyi paraméterek elvileg bármilyen pontossággal meghatározhatók.

• Két szinuszos jellel vizsgálva a modellt, a fent elmondottak továbbra is érvényesek, ha a frekvenciák között több kritikus sáv távolság van. Ha a két jel ugyanabba a részsávba esik, akkor $\Omega_k(t)$ átlaga a két frekvencia számtani átlaga.

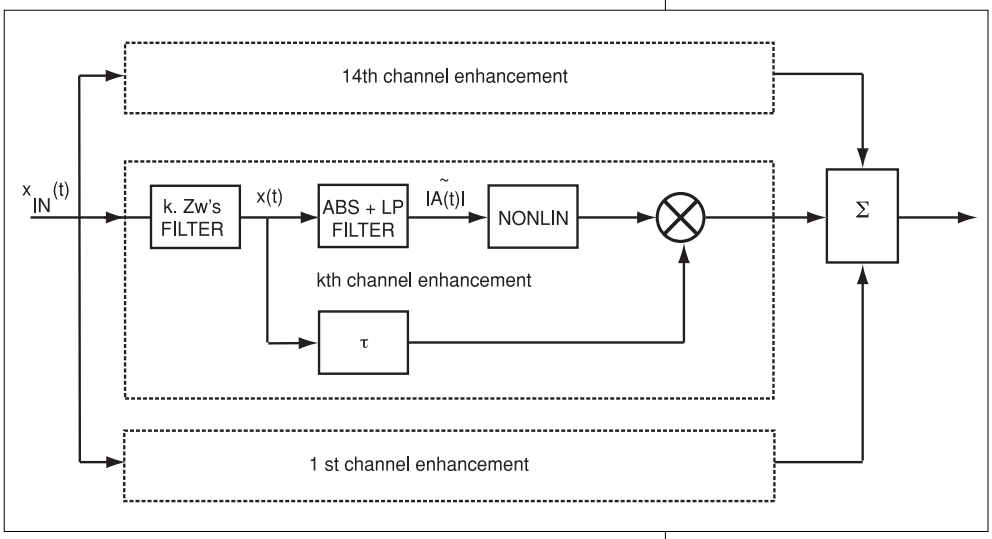
• A modell felhasználásával meghatározható a virtuális hang, ugyanis a hiányzó alaphang feletti részsávok jeléből $\Omega_i(t)$ és $\Omega_k(t)$ meghatározható a váltóáramú komponensek periódusideje, melyek a hiányzó alaphang pe-

riódusidejével egyeznek meg. Az így meghatározott periódusidőt nem lehet befolyásolni a hiányzó alaphang helyére beadott keskenysávú zajjal, és nem lehetséges a hiányzó alaphang környezetében lebegést elérni. Ha azonban a jel az alapot tartalmazza, akkor lebegés jön létre, hiszen a szóban forgó részsáv pillanatnyi paraméterei pontosan követik a lebegést.

• A modellen a bizonytalansági reláció csak meglehetősen bonyolultan értelmezhető. Azonban megmutatható, hogy a modell felhasználásával kisebb felbontás is elérhető, mint a hallásra publikált érték [7].

6. Beszéd kiemelése háttérzajból

A zajos háttérből történő beszédkiemelésre a hallásmodell egyszerűsített változatát célszerű használni. Az eljárás blokkvázlatát az *5. ábra* mutatja.



5. ábra Zajos háttérből történő beszédkiemelés egyszerűsített blokkvázlata

A megoldás elméleti alapját az úgynevezett „optimumszűrő” szolgáltatja [11]. Ha egy rendszer bemenetére zajos jel kerül, azaz

$$x(t) = s(t) + n(t), \tag{3}$$

akkor a rendszer jellemzőit úgy célszerű megválasztani, hogy a kimeneti y(t) jel minél többet tartalmazzon a hasznos jelből, és minél kevesebbet a zavaró jelből. A feladat általános esetben nem oldható meg, ezért a feladatot célszerű lineáris rendszerre korlátozni. Az így nyert lineáris rendszer az „optimumszűrő”.

A Wiener-Hopf integrálegyenlet megoldása adja a keresett rendszer K(ω) átviteli karakterisztikáját, amely a jel és a zaj teljesítmény sűrűség spektrumával kifejezve a következő:

$$K(\omega) = \frac{g_S(\omega)}{g_S(\omega) + g_N(\omega)}. \tag{4}$$

Ennek alapján a k-adik csatornában az $A_k^*(t)$ súlytényezőt a következő értékre kell beállítani:

$$A_k(t) = K(\omega_k) = \frac{g_S(\omega_k)}{g_S(\omega_k) + g_N(\omega_k)}. \tag{5}$$

Ebben az egyenletben a teljesítmény sűrűségspektrumok nem ismertek, de igen jó becslések adhatók értékükre.

Abból a felismerésből kiindulva, hogy a beszéd mindig tartalmaz szüneteket, következik, hogy $A_k^L(t)$ minimumának négyzete arányos a háttérzaj teljesítményével, azaz

$$g_N(\omega_k) = c\left(\min\left\{A_k^L(t)\right\}\right)^2, \tag{6}$$

továbbá $A_k^L(t)$ pillanatnyi értékének négyzete arányos a jel és a zaj teljesítményének az összegével, hiszen a jel és a zaj kölcsönösen független folyamatok. Mindezek alapján írhatjuk, hogy

$$g_S(\omega_k) + g_N(\omega_k) = c\left(A_k^L(t)\right)^2. \tag{7}$$

Normalizáljuk $A_k^L(t)$ értékét a minimumával, és vezessük be a következő egyszerűsítő jelölést:

$$z_t = \frac{A_k^L(t)}{\min\left\{A_k^L(t)\right\}}. \tag{8}$$

Fentieket felhasználva néhány egyszerű átalakítás után azt kapjuk, hogy a súlytényező értéke:

$$A_k(t) = \frac{z_t^2 - 1}{z_t^2}. \tag{9}$$

Ez a lassan változó jel (az 5. ábra közepén $|\tilde{A}(t)|$-vel jelölve) kerül a (9) által meghatározott nemlineáris karakterisztikára, melynek kimeneti jele állítja be minden egyes részsáv erősítését.

Természetesen, ha a beszéd nem tartalmaz háttérzajt, akkor a csatorna jele változatlanul kerül az összegzőre, hiszen ebben az esetben minden súlytényező értéke $A_k(t) = 1$.

7. Záró megjegyzések

A) A 6. pontban ismertetett eljárás igen jól használható régi zajos felvételek, hanglemezek tűzörejének, valamint régi filmek hanganyagának tisztítására. Ezekben az esetekben az optimálisan megtisztított jel hangzása nem a legkellemesebb, ezért a normalizálást nem a minimummal, hanem egy kisebb értékkel célszerű elvégezni. Természetesen így kevesebb háttérzaj kerül eltávolításra, de kellemesebbnek halljuk a megtisztított anyagot. A legjobb megoldásnak azt tartjuk, ha a helyes arány beállítását hangmérnök végzi.

B) Ha a háttérzaj egészen speciális (pl. egy üzemcsarnokban a beszédnél is hangosabb csattanások), az eljárás természetesen csak a beszédszünetekben hallható zajt csökkenti, a csattanásokat nem, hiszen a háttérzajból éppen úgy kiemelkedik, mint a beszéd. Ilyen speciális zavarok csökkentéséhez további kiegészítésekre van szükség. Például pitch detektor felhasználásával a beszéd maximumai meghatározhatók. A csattanások szintje egy, a zajforráshoz közel elhelyezett mikrofon segítségével csökkenthető (ismert jel elnyomás). Ez a megoldás bármilyen típusú zaj esetén használható, ha háttérzaj jól definiálható forrásból származik. Ha ez nem áll fenn, akkor ez a megoldás igen kevés eredménnyel kecsegtet. Például egy autóban a szélvédő bal és jobb oldalánál elhelyezett mikrofonok jelei gyakorlatilag függetlenek, ezért egyik jel sem csökkenthető a másik felhasználásával.

C) Ha a háttérzaj összemérhető a beszéd hangosságával, azaz a jel-zaj viszony kisebb 10 dB-nél, akkor előfordulhat, hogy a tisztított beszéd rosszabbul érthető, mint az eredeti. Ilyen esetben a tisztítás hatásfokát csökkenteni kell. Ez a feladat megoldható, ha a minimumokon kívül a maximumokat is figyeljük, és megpróbálunk a jel-zaj viszonyra becslést adni. Sajnos minden típusra más-más algoritmust kellene kidolgozni (pl. a jelből kiemelkedő csattanás ad egy maximumot, de ennek semmi köze sincs a jel-zaj viszonyhoz).

D) A nemlineáris hallásmodell [10] felhasználásával és további kiegészítésével majdnem minden feladat megoldását sikerült szimulálni. Rendkívül jól használható ismert jel elnyomásra, források szétválasztására, visszhang csökkentésére, zöngés-zöngétlen döntő, valamint pitch detektor készítésére, továbbá beszéd tömörítésére [12, 13]. A minőség még tökéletes (az eredetitől megkülönböztethetetlen) maradt 1000 bit/s átviteli sebesség esetén is (késleltetési idő kb. 100 ms). Kisebb sebességű átvitel is megvalósítható, de természetesen csak minőségromlás és a késleltetési idő további növekedése árán.

Köszönetnyilvánítás

A hallásmodell elméleti hátterének kidolgozásában nyújtott segítségért ezúton is szeretnénk köszönetet mondani dr. Papp Lászlónak és dr. Osváth Lászlónak.

Irodalom

[1] G. S. Ohm:
Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen.
Ann. der Phisik, Vol. 59, No. 8, 512-565, 1843.

[2] H.v. Helmholtz:
Die Lehre von den Tonempfindungen.
Braunschweig, 1863, 1913.

[3] D. Gabor:
Acoustical Quanta and the Theory of Hearing.
Nature, Vol. 159, 591-692, 1947.

[4] Békésy, Gy., Rosenblith, W. A.:
The early history of hearing observations and theories.
J. Acoust. Soc. Am. Vol. 20, 1948.

[5] H. Fletcher:
Speech and Hearing.
Nostrand C., New York, 1950.

[6] E. Zwicker, R. Feldtkeller:
Das Ohr als Nachrichtenempfnger.
Hirzel V., Stuttgart, 1967.

[7] L. M. Grobben:
Appreciation of Short Tones.
Seventh International Congress on Acoustics, Budapest, Vol. 3, 329-332, 1971.

[8] R. Földvári:
Adaptive Sampling.
Periodica Politechnica Electrical Engineering, Vol. 33, No. 3, Budapest, 1989.

[9] Tarnóczy T.:
Einführung in die musikalische Akustik.
Akadémiai Kiadó, Budapest, 1991.

[10] R. Földvári:
Generalized instantaneous amplitude and frequency functions and their application for pitch frequency determination.
Journal of Circuits, Systems, and Computers, Vol. 5, No. 2, 1995.

[11] R. Földvári, Gy. Ács:
Speech Enhancement Based on a New Hearing Model.
19th Czech-Hungarian-Polish Workshop on Circuit Theory and Applications, Prague, 1966.

[12] R. Földvári, Gy. Ács:
Speech and Music Coder Based on a New Hearing Model.
7th Conference and Exhibition on Television and Sound Technique, Budapest, 1996.

[13] R. Földvári, L. Gyimesi:
Very Low Bit Rate Voice Coder Based on a Nonlinear Hearing Model.
Eurospeech '99 – 6th European Conference of Speech Communication and Technology, Budapest, 1999.